

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2001-117597

(43)Date of publication of application : 27.04.2001

(51)Int.Cl. G10L 21/04
G10K 15/04
G10L 15/00
G10L 13/06

(21)Application number : 11-300268

(71)Applicant : YAMAHA CORP
UNIV POMPEU FABRA

(22)Date of filing : 21.10.1999

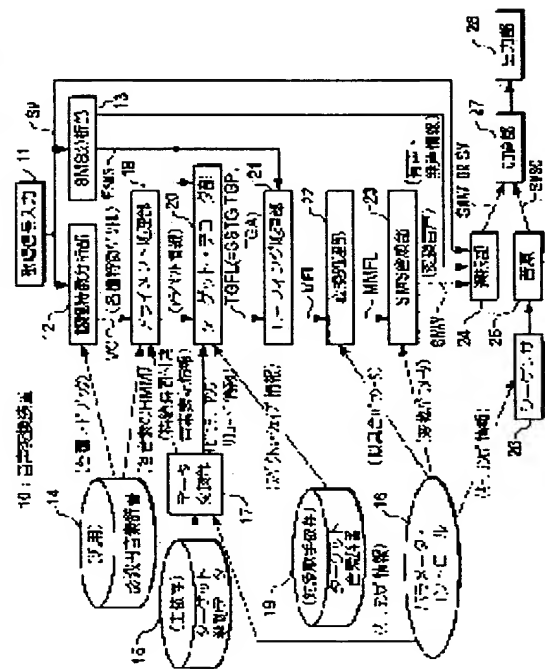
(72)Inventor : YOSHIOKA YASUO
XAVIER SERA
MARC SHIIMENTSU
JORDI BONADA

(54) DEVICE AND METHOD FOR VOICE CONVERSION AND METHOD OF GENERATING DICTIONARY FOR VOICE CONVERSION

(57)Abstract:

PROBLEM TO BE SOLVED: To make the inputted voice of a singer possible to resemble the singing of a target singer and to reduce the capacity of analytic data of the target singer.

SOLUTION: Input frame data FSMS corresponding to an input voice signal SV are extracted and alignment adjustment for synchronizing the input frame data FSMS and target frame data TGFL to be generated is performed, and the target frame data TGFL are generated on the basis of target frame generation data previously extracted from a target voice and on the basis of the input frame data FSMS and target frame data TGFL, a converted sound signal is generated and outputted.



BEST AVAILABLE COPY

LEGAL STATUS

[Date of request for examination]

23.02.2006

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

THIS PAGE BLANK (USPTO)

[Date of extinction of right]

THIS PAGE BLANK (USPTO)

【特許請求の範囲】

【請求項1】 入力音声信号に対応する入力フレームデータを抽出する入力フレームデータ抽出手段と、前記入力フレームデータと生成すべきターゲットフレームデータとの間の同期をとるべくアライメント調整を行うとともに、ターゲット音声から予め抽出したターゲットフレーム生成データに基づいて前記ターゲットフレームデータを生成するターゲットフレームデータ生成手段と、前記入力フレームデータ及び前記ターゲットフレームデータに基づいて変換音声信号を生成し出力する変換音声信号生成手段と、を備えたことを特徴とする音声変換装置。

【請求項2】 請求項1記載の音声変換装置において、前記ターゲットフレーム生成データには、ターゲット音声のピッチ軌跡データ、接続時間付き音素表情報および音素毎のスペクトル・シェイプが含まれることを特徴とする音声変換装置。

【請求項3】 請求項1または請求項2記載の音声変換装置において、前記ターゲットフレーム生成情報には、アンブリチュード軌跡情報が含まれることを特徴とする音声変換装置。

【請求項4】 請求項1記載の音声変換装置において、前記ターゲットフレームデータ生成手段は、前記入力音声信号から特徴ベクトルを抽出する特徴ベクトル抽出手段と、

予め記憶した認識用音素データ及び予め記憶したターゲット挙動データに基づいて前記入力フレームデータと前記ターゲットフレームデータとの間の時間的位置関係を判別し、前記時間的位置関係に対応するアライメントデータを出力するアライメント処理手段と、前記ターゲットデータ、前記アライメントデータ、予め記憶したターゲット音素データ及び前記入力フレームデータに基づいて前記ターゲットフレームデータを生成するターゲットデコード手段と、を備えたことを特徴とする音声変換装置。

【請求項5】 請求項4記載の音声変換装置において、予めターゲット音声の挙動を表すターゲット挙動データを記憶するターゲット挙動データ記憶手段と、前記ターゲット挙動データおよび外部より入力されたパラメータコントロールデータに基づいてピッチデータ、アンブリチュードデータおよび接続時間付き音素表記データを生成し出力するデータ変換手段と、を備えたことを特徴とする音声変換装置。

【請求項6】 請求項4記載の音声変換装置において、前記ターゲットデコード手段は、ターゲット音素辞書内の同一音素についての二つのピッチに対応したスペクトル・シェイプを用いて補間処理を行い、所望のピッチに対応するスペクトル・シェイプを算出するスペクトル・シェイプ補間手段を備えたことを特徴とする音声変換装

置。

【請求項7】 請求項6記載の音声変換装置において、音素が安定状態あるいは遷移状態のいずれにあるかを判別する状態判別手段を備え、

前記スペクトル・シェイプ補間手段は、前記状態判別手段の判別状態に基づいて前記音素が遷移状態にある場合に、遷移元の音素に対応するスペクトル・シェイプ及び遷移先の音素に対応するスペクトル・シェイプを用いて前記補間処理を行うことを特徴とする音声変換装置。

【請求項8】 請求項6記載の音声変換装置において、前記スペクトル・シェイプ補間手段は、二つのスペクトル・シェイプを用いて補間を行うに際し、前記二つのスペクトル・シェイプ間におけるスペクトル・シェイプ遷移関数を用いて前記補間処理を行うことを特徴とする音声変換装置。

【請求項9】 請求項8記載の音声変換装置において、前記遷移関数は、線形関数あるいは非線形関数として予め定義されていることを特徴とする音声変換装置。

【請求項10】 請求項8記載の音声変換装置において、

前記二つのスペクトル・シェイプを周波数軸上でそれぞれ複数の領域に分け、各領域毎に前記遷移関数を定めることを特徴とする音声変換装置。

【請求項11】 請求項8記載の音声変換装置において、前記スペクトル・シェイプ補間手段は、前記遷移先の音素に対応させて前記遷移関数を定めることを特徴とする音声変換装置。

【請求項12】 請求項8記載の音声変換装置において、

前記スペクトル・シェイプ補間手段は、前記入力音声信号に対応する、あるいは、前記ターゲット音声に対応するピッチ若しくはスペクトル・シェイプに基づいて前記遷移関数をリアルタイムで決定する遷移関数リアルタイム決定手段と、を備えたことを特徴とする音声変換装置。

【請求項13】 請求項8記載の音声変換装置において、

前記スペクトル・シェイプ補間手段は、前記二つのスペクトル・シェイプを周波数軸上でそれぞれ複数の領域に分け、各領域に属する前記二つのスペクトル・シェイプ上の実在の周波数およびマグニチュードの組に対し、前記遷移関数としての線形関数を用いた前記補間処理を前記複数の領域にわたって行うことを特徴とする音声変換装置。

【請求項14】 請求項13記載の音声変換装置において、

前記スペクトル・シェイプ補間手段は、前記各領域に属する一方のスペクトル・シェイプの周波数である第1周波数及び当該第1周波数に対応する他方のスペクトル・シェイプの周波数である第2周波数を前記線形関数を用

いて補間することにより補間周波数を算出する周波数補間手段と、

前記各領域に属する一方のスペクトル・シェイプのマグニチュードである第1マグニチュードおよび当該第1マグニチュードに対応する他方のスペクトル・シェイプのマグニチュードである第2マグニチュードを前記線形関数を用いて補間するマグニチュード補間手段と、
を備えたことを特徴とする音声変換装置。

【請求項15】 請求項4記載の音声変換装置において、
前記ターゲットフレームデータは、ターゲットスペクトル・シェイプを含み、
前記ターゲットデコード手段は、前記ターゲットスペクトル・シェイプのアンプリチュードに応じて当該ターゲットスペクトル・シェイプのスペクトル傾きを補正するスペクトル傾き補正手段を備えたことを特徴とする音声変換装置。

【請求項16】 請求項15記載の音声変換装置において、
前記スペクトル傾き補正手段は、前記ターゲットスペクトル・シェイプのスペクトル傾きと、当該ターゲットスペクトル・シェイプに対応し、かつ、前記入力音声信号に対応する入力スペクトル・シェイプのスペクトル傾きと、の差に相当するスペクトル傾き補正フィルタ手段を備えたことを特徴とする音声変換装置。

【請求項17】 入力音声信号に対応する入力フレームデータを抽出する入力フレームデータ抽出過程と、
前記入力フレームデータと生成すべきターゲットフレームデータとの間の同期をとるべくアライメント調整を行うとともに、ターゲット音声から予め抽出したターゲットフレーム生成データに基づいて前記ターゲットフレームデータを生成するターゲットフレームデータ生成過程と、

前記入力フレームデータ及び前記ターゲットフレームデータに基づいて変換音声信号を生成し出力する変換音声信号生成過程と、

を備えたことを特徴とする音声変換方法。

【請求項18】 請求項17記載の音声変換方法において、
前記ターゲットフレームデータ生成過程は、前記入力音声信号から特徴ベクトルを抽出する特徴ベクトル抽出過程と、
予め記憶した認識用音素データ及び予め記憶したターゲット挙動データに基づいて前記入力フレームデータと前記ターゲットフレームデータとの間の時間的位置関係を判別し、前記時間的位置関係に対応するアライメントデータを出力するアライメント処理過程と、

前記ターゲットデータ、前記アライメントデータ、予め記憶したターゲット音素データ及び前記入力フレームデータに基づいて前記ターゲットフレームデータを生成す

るターゲットデコード過程と、
を備えたことを特徴とする音声変換方法。

【請求項19】 請求項18記載の音声変換方法において、
予め記憶したターゲット音声の挙動を表すターゲット挙動データおよび外部より入力されたパラメータコントロールデータに基づいてピッチデータ、アンプリチュードデータおよび接続時間付き音素表記データを生成し出力するデータ変換過程と、
を備えたことを特徴とする音声変換方法。

【請求項20】 請求項18記載の音声変換方法において、
前記ターゲットデコード過程は、ターゲット音素辞書内の同一音素についての二つのピッチに対応したスペクトル・シェイプを用いて補間処理を行い、所望のピッチに対応するスペクトル・シェイプを算出するスペクトル・シェイプ補間過程を備えたことを特徴とする音声変換方法。

【請求項21】 請求項20記載の音声変換方法において、
音素が安定状態あるいは遷移状態のいずれにあるかを判別する状態判別過程を備え、
前記スペクトル・シェイプ補間過程は、前記状態判別過程の判別状態に基づいて前記音素が遷移状態にある場合に、遷移元の音素に対応するスペクトル・シェイプ及び遷移先の音素に対応するスペクトル・シェイプを用いて前記補間処理を行うことを特徴とする音声変換方法。

【請求項22】 請求項20記載の音声変換方法において、
前記スペクトル・シェイプ補間過程は、二つのスペクトル・シェイプを用いて補間を行うに際し、前記二つのスペクトル・シェイプ間におけるスペクトル・シェイプ遷移関数を用いて前記補間処理を行うことを特徴とする音声変換方法。

【請求項23】 請求項22記載の音声変換方法において、
前記遷移関数は、線形関数あるいは非線形関数として予め定義されていることを特徴とする音声変換方法。

【請求項24】 請求項22記載の音声変換方法において、
前記二つのスペクトル・シェイプを周波数軸上でそれぞれ複数の領域に分け、各領域毎に前記遷移関数を定めることを特徴とする音声変換方法。

【請求項25】 請求項22記載の音声変換方法において、
前記スペクトル・シェイプ補間過程は、前記遷移先の音素に対応させて前記遷移関数を定めることを特徴とする音声変換方法。

【請求項26】 請求項22記載の音声変換方法において、

前記スペクトル・シェイプ補間過程は、前記入力音声信号に対応する、あるいは、前記ターゲット音声に対応するピッチ若しくはスペクトル・シェイプに基づいて前記遷移関数をリアルタイムで決定する遷移関数リアルタイム決定過程と、

を備えたことを特徴とする音声変換方法。

【請求項27】 請求項22記載の音声変換方法において、

前記スペクトル・シェイプ補間過程は、前記二つのスペクトル・シェイプを周波数軸上でそれぞれ複数の領域に分け、各領域に属する前記二つのスペクトル・シェイプ上の実在の周波数およびマグニチュードの組に対し、前記遷移関数としての線形関数を用いた前記補間処理を前記複数の領域にわたって行うことを特徴とする音声変換方法。

【請求項28】 請求項27記載の音声変換方法において、

前記スペクトル・シェイプ補間過程は、前記各領域に属する一方のスペクトル・シェイプの周波数である第1周波数及び当該第1周波数に対応する他方のスペクトル・シェイプの周波数である第2周波数を前記線形関数を用いて補間することにより補間周波数を算出する周波数補間過程と、

前記各領域に属する一方のスペクトル・シェイプのマグニチュードである第1マグニチュードおよび当該第1マグニチュードに対応する他方のスペクトル・シェイプのマグニチュードである第2マグニチュードを前記線形関数を用いて補間するマグニチュード補間過程と、
を備えたことを特徴とする音声変換方法。

【請求項29】 請求項18記載の音声変換方法において、

前記ターゲットフレームデータは、ターゲットスペクトル・シェイプを含み、

前記ターゲットデコード過程は、前記ターゲットスペクトル・シェイプのアンプリチュードに応じて当該ターゲットスペクトル・シェイプのスペクトル傾きを補正するスペクトル傾き補正過程を備えたことを特徴とする音声変換方法。

【請求項30】 請求項29記載の音声変換方法において、

前記スペクトル傾き補正過程は、前記ターゲットスペクトル・シェイプのスペクトル傾きと、当該ターゲットスペクトル・シェイプに対応し、かつ、前記入力音声信号に対応する入力スペクトル・シェイプのスペクトル傾きと、の差に相当するスペクトル傾き補正を行うことを特徴とする音声変換方法。

【請求項31】 ターゲット歌手の音声であるターゲット音声に対応する音声変換用辞書を生成する音声変換用辞書の生成方法において、

各音素について前記ターゲット歌手により低域側ピッチ

から高域側ピッチにわたって連続的に発声させた音素発声信号に基づいてスペクトル・シェイプを算出し、
得られたスペクトル・シェイプを同一のスペクトル・シェイプとみなせるスペクトル・シェイプが属するピッチの区間をセグメントと定義し、

前記セグメントに属する全ての前記スペクトル・シェイプを予め定めた所定の周波数帯域幅を有する周波数帯域毎に分割し、

各前記周波数帯域の中心周波数を算出し、

各前記周波数帯域に属するスペクトル・シェイプの当該周波数帯域に対応する前記中心周波数におけるマグニチュードの平均値を平均マグニチュードとし、

前記平均マグニチュードを前記セグメントに属する全ての前記周波数帯域について算出し、

前記セグメントに属する全ての前記周波数帯域について算出した平均マグニチュードを前記セグメント区間の平均ピッチにおけるターゲットスペクトル・シェイプとして、前記音声変換用辞書に格納する、

ことを特徴とする音声変換用辞書の生成方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】この発明は、処理対象となる音声を目標とする他の音声に近似させる音声変換装置、音声変換方法ならびに音声変換を行うに際し用いる他の音声に対応する音声変換用辞書を生成する音声変換用辞書の生成方法に係り、特にカラオケ装置に用いるのに好適な音声変換装置、音声変換方法及び音声変換用辞書の生成方法に関する。

【0002】

【従来の技術】入力された音声の周波数特性などを変えて出力する音声変換装置は種々開発されており、例えば、カラオケ装置の中には、歌手の歌った歌声のピッチを変換して、男性の声を女性の声に、あるいはその逆に変換させるものもある（例えば、特表平8-508581号）。

【0003】

【発明が解決しようとする課題】しかしながら、従来の音声変換装置においては、音声の変換（例えば、男声→女声、女声→男声など）は行われるものの、単に声質を変えるだけに止まっていたので、例えば、特定の歌唱者（例えば、プロの歌手）の声に似せるように変換するということはできなかった。また、声質だけでなく、歌い方までも特定の歌唱者に似させるという、ものまねのような機能があれば、カラオケ装置などにおいては大変に面白いが、従来の音声変換装置ではこのような処理は不可能であった。

【0004】そこで、発明者らは、声質を目標（ターゲット）とする歌唱者（ターゲット歌唱者）の声に似させるために、ターゲット歌唱者の音声进行分析し、得られた分析データである正弦波成分属性ピッチ、アンプリチュ

ード、スペクトル・シェイプ及び残差成分を1曲分全てのフレームについてターゲットフレームデータとして保持し、入力音声进行分析して得られる入力ターゲットフレームデータとの同期をとって、変換処理を行うことによりターゲット歌唱者の声に似せるように変換を行う音声変換装置を提案している(特願平10-183338号等参照)。上記音声変換装置は、声質だけでなく、歌い方までも特定の歌唱者に似させることができるが、ターゲット歌唱者の分析データが一曲毎に必要となり、複数の曲の分析データを記憶させるような場合には、データ量が膨大になってしまうという不具合があった。

【0005】そこで、本発明の目的は、入力された歌唱者の音声を目標とする歌唱者の歌い方に似せることができるとともに、ターゲット歌唱者の分析データの容量を低減することが可能な音声変換装置、音声変換方法および音声変換用辞書の生成方法を提供することにある。

【0006】

【課題を解決するための手段】上記課題を解決するため、請求項1記載の構成は、入力音声信号に対応する入力フレームデータを抽出する入力フレームデータ抽出手段と、前記入力フレームデータと生成すべきターゲットフレームデータとの間の同期をとるべくアライメント調整を行うとともに、ターゲット音声から予め抽出したターゲットフレーム生成データに基づいて前記ターゲットフレームデータを生成するターゲットフレームデータ生成手段と、前記入力フレームデータ及び前記ターゲットフレームデータに基づいて変換音声信号を生成し出力する変換音声信号生成手段と、を備えたことを特徴としている。

【0007】請求項2記載の構成は、請求項1記載の構成において、前記ターゲットフレーム生成データには、ターゲット音声のピッチ軌跡データ、接続時間付き音素表記情報および音素毎のスペクトル・シェイプが含まれることを特徴としている。

【0008】請求項3記載の構成は、請求項1または請求項2記載の構成において、前記ターゲットフレーム生成情報には、アンプリチュード軌跡情報が含まれることを特徴としている。

【0009】請求項4記載の構成は、請求項1記載の構成において、前記ターゲットフレームデータ生成手段は、前記入力音声信号から特徴ベクトルを抽出する特徴ベクトル抽出手段と、予め記憶した認識用音素データ及び予め記憶したターゲット挙動データに基づいて前記入力フレームデータと前記ターゲットフレームデータとの間の時間的位置関係を判別し、前記時間的位置関係に対応するアライメントデータを出力するアライメント処理手段と、前記ターゲットデータ、前記アライメントデータ、予め記憶したターゲット音素データ及び前記入力フレームデータに基づいて前記ターゲットフレームデータを生成するターゲットデコード手段と、を備えたことを

特徴としている。

【0010】請求項5記載の構成は、請求項4記載の構成において、予めターゲット音声の挙動を表すターゲット挙動データを記憶するターゲット挙動データ記憶手段と、前記ターゲット挙動データおよび外部より入力されたパラメータコントロールデータに基づいてピッチデータ、アンプリチュードデータおよび接続時間付き音素表記データを生成し出力するデータ変換手段と、を備えたことを特徴としている。

【0011】請求項6記載の構成は、請求項4記載の構成において、前記ターゲットデコード手段は、ターゲット音素辞書内の同一音素についての二つのピッチに対応したスペクトル・シェイプを用いて補間処理を行い、所望のピッチに対応するスペクトル・シェイプを算出するスペクトル・シェイプ補間手段を備えたことを特徴としている。

【0012】請求項7記載の構成は、請求項6記載の構成において、音素が安定状態あるいは遷移状態のいずれにあるかを判別する状態判別手段を備え、前記スペクトル・シェイプ補間手段は、前記状態判別手段の判別状態に基づいて前記音素が遷移状態にある場合に、遷移元の音素に対応するスペクトル・シェイプ及び遷移先の音素に対応するスペクトル・シェイプを用いて前記補間処理を行うことを特徴としている。

【0013】請求項8記載の構成は、請求項6記載の構成において、前記スペクトル・シェイプ補間手段は、二つのスペクトル・シェイプを用いて補間を行うに際し、前記二つのスペクトル・シェイプ間におけるスペクトル・シェイプ遷移関数を用いて前記補間処理を行うことを特徴としている。

【0014】請求項9記載の構成は、請求項8記載の構成において、前記遷移関数は、線形関数あるいは非線形関数として予め定義されていることを特徴としている。

【0015】請求項10記載の構成は、請求項8記載の構成において、前記二つのスペクトル・シェイプを周波数軸上でそれぞれ複数の領域に分け、各領域毎に前記遷移関数を定めることを特徴としている。

【0016】請求項11記載の構成は、請求項8記載の構成において、前記スペクトル・シェイプ補間手段は、前記遷移先の音素に対応させて前記遷移関数を定めることを特徴としている。

【0017】請求項12記載の構成は、請求項8記載の構成において、前記スペクトル・シェイプ補間手段は、前記入力音声信号に対応する、あるいは、前記ターゲット音声に対応するピッチ若しくはスペクトル・シェイプに基づいて前記遷移関数をリアルタイムで決定する遷移関数リアルタイム決定手段と、を備えたことを特徴としている。

【0018】請求項13記載の構成は、請求項8記載の構成において、前記スペクトル・シェイプ補間手段は、

前記二つのスペクトル・シェイプを周波数軸上でそれぞれ複数の領域に分け、各領域に属する前記二つのスペクトル・シェイプ上の実在の周波数およびマグニチュードの組に対し、前記遷移関数としての線形関数を用いた前記補間処理を前記複数の領域にわたって行うことを特徴としている。

【0019】請求項14記載の構成は、請求項13記載の構成において、前記スペクトル・シェイプ補間手段は、前記各領域に属する一方のスペクトル・シェイプの周波数である第1周波数及び当該第1周波数に対応する他方のスペクトル・シェイプの周波数である第2周波数を前記線形関数を用いて補間することにより補間周波数を算出する周波数補間手段と、前記各領域に属する一方のスペクトル・シェイプのマグニチュードである第1マグニチュードおよび当該第1マグニチュードに対応する他方のスペクトル・シェイプのマグニチュードである第2マグニチュードを前記線形関数を用いて補間するマグニチュード補間手段と、を備えたことを特徴としている。

【0020】請求項15記載の構成は、請求項4記載の構成において、前記ターゲットフレームデータは、ターゲットスペクトル・シェイプを含み、前記ターゲットデコード手段は、前記ターゲットスペクトル・シェイプのアンプリチュードに応じて当該ターゲットスペクトル・シェイプのスペクトル傾きを補正するスペクトル傾き補正手段を備えたことを特徴としている。

【0021】請求項16記載の構成は、請求項15記載の構成において、前記スペクトル傾き補正手段は、前記ターゲットスペクトル・シェイプのスペクトル傾きと、当該ターゲットスペクトル・シェイプに対応し、かつ、前記入力音声信号に対応する入力スペクトル・シェイプのスペクトル傾きと、の差に相当するスペクトル傾き補正フィルタ手段を備えたことを特徴としている。

【0022】請求項17記載の構成は、入力音声信号に対応する入力フレームデータを抽出する入力フレームデータ抽出過程と、前記入力フレームデータと生成すべきターゲットフレームデータとの間の同期をとるべくアライメント調整を行うとともに、ターゲット音声から予め抽出したターゲットフレーム生成データに基づいて前記ターゲットフレームデータを生成するターゲットフレームデータ生成過程と、前記入力フレームデータ及び前記ターゲットフレームデータに基づいて変換音声信号を生成し出力する変換音声信号生成過程と、を備えたことを特徴としている。

【0023】請求項18記載の構成は、請求項17記載の構成において、前記ターゲットフレームデータ生成過程は、前記入力音声信号から特徴ベクトルを抽出する特徴ベクトル抽出過程と、予め記憶した認識用音素データ及び予め記憶したターゲット挙動データに基づいて前記入力フレームデータと前記ターゲットフレームデータと

の間の時間的位置関係を判別し、前記時間的位置関係に対応するアライメントデータを出力するアライメント処理過程と、前記ターゲットデータ、前記アライメントデータ、予め記憶したターゲット音素データ及び前記入力フレームデータに基づいて前記ターゲットフレームデータを生成するターゲットデコード過程と、を備えたことを特徴としている。

【0024】請求項19記載の構成は、請求項18記載の構成において、予め記憶したターゲット音声の挙動を表すターゲット挙動データおよび外部より入力されたパラメータコントロールデータに基づいてピッチデータ、アンプリチュードデータおよび接続時間付き音素表記データを生成し出力するデータ変換過程と、を備えたことを特徴としている。

【0025】請求項20記載の構成は、請求項18記載の構成において、前記ターゲットデコード過程は、ターゲット音素辞書内の同一音素についての二つのピッチに対応したスペクトル・シェイプを用いて補間処理を行い、所望のピッチに対応するスペクトル・シェイプを算出するスペクトル・シェイプ補間過程を備えたことを特徴としている。

【0026】請求項21記載の構成は、請求項20記載の構成において、音素が安定状態あるいは遷移状態のいずれにあるかを判別する状態判別過程を備え、前記スペクトル・シェイプ補間過程は、前記状態判別過程の判別状態に基づいて前記音素が遷移状態にある場合に、遷移元の音素に対応するスペクトル・シェイプ及び遷移先の音素に対応するスペクトル・シェイプを用いて前記補間処理を行うことを特徴としている。

【0027】請求項22記載の構成は、請求項20記載の構成において、前記スペクトル・シェイプ補間過程は、二つのスペクトル・シェイプを用いて補間を行うに際し、前記二つのスペクトル・シェイプ間におけるスペクトル・シェイプ遷移関数を用いて前記補間処理を行うことを特徴としている。

【0028】請求項23記載の構成は、請求項22記載の構成において、前記遷移関数は、線形関数あるいは非線形関数として予め定義されていることを特徴としている。

【0029】請求項24記載の構成は、請求項22記載の構成において、前記二つのスペクトル・シェイプを周波数軸上でそれぞれ複数の領域に分け、各領域毎に前記遷移関数を定めることを特徴としている。

【0030】請求項25記載の構成は、請求項22記載の構成において、前記スペクトル・シェイプ補間過程は、前記遷移先の音素に対応させて前記遷移関数を定めることを特徴としている。

【0031】請求項26記載の構成は、請求項22記載の構成において、前記スペクトル・シェイプ補間過程は、前記入力音声信号に対応する、あるいは、前記ター

ゲット音声に対応するピッチ若しくはスペクトル・シェイプに基づいて前記遷移関数をリアルタイムで決定する遷移関数リアルタイム決定過程と、を備えたことを特徴としている。

【0032】請求項27記載の構成は、請求項22記載の構成において、前記スペクトル・シェイプ補間過程は、前記二つのスペクトル・シェイプを周波数軸上でそれぞれ複数の領域に分け、各領域に属する前記二つのスペクトル・シェイプ上の実在の周波数およびマグニチュードの組に対し、前記遷移関数としての線形関数を用いた前記補間処理を前記複数の領域にわたって行うことを特徴としている。

【0033】請求項28記載の構成は、請求項27記載の構成において、前記スペクトル・シェイプ補間過程は、前記各領域に属する一方のスペクトル・シェイプの周波数である第1周波数及び当該第1周波数に対応する他方のスペクトル・シェイプの周波数である第2周波数を前記線形関数を用いて補間することにより補間周波数を算出する周波数補間過程と、前記各領域に属する一方のスペクトル・シェイプのマグニチュードである第1マグニチュードおよび当該第1マグニチュードに対応する他方のスペクトル・シェイプのマグニチュードである第2マグニチュードを前記線形関数を用いて補間するマグニチュード補間過程と、を備えたことを特徴としている。

【0034】請求項29記載の構成は、請求項18記載の構成において、前記ターゲットフレームデータは、ターゲットスペクトル・シェイプを含み、前記ターゲットデコード過程は、前記ターゲットスペクトル・シェイプのアンプリチュードに応じて当該ターゲットスペクトル・シェイプのスペクトル傾きを補正するスペクトル傾き補正過程を備えたことを特徴としている。

【0035】請求項30記載の構成は、請求項29記載の構成において、前記スペクトル傾き補正過程は、前記ターゲットスペクトル・シェイプのスペクトル傾きと、当該ターゲットスペクトル・シェイプに対応し、かつ、前記入力音声信号に対応する入力スペクトル・シェイプのスペクトル傾きと、の差に相当するスペクトル傾き補正を行うことを特徴としている。

【0036】請求項31記載の構成は、ターゲット歌手の音声であるターゲット音声に対応する音声変換用辞書を生成する音声変換用辞書の生成方法において、各音素について前記ターゲット歌手により低域側ピッチから高域側ピッチにわたって連続的に発声させた音素発声信号に基づいてスペクトル・シェイプを算出し、得られたスペクトル・シェイプを同一のスペクトル・シェイプとみなせるスペクトル・シェイプが属するピッチの区間をセグメントと定義し、前記セグメントに属する全ての前記スペクトル・シェイプを予め定めた所定の周波数帯域幅を有する周波数帯域毎に分割し、各前記周波数帯域の中

心周波数を算出し、各前記周波数帯域に属するスペクトル・シェイプの当該周波数帯域に対応する前記中心周波数におけるマグニチュードの平均値を平均マグニチュードとし、前記平均マグニチュードを前記セグメントに属する全ての前記周波数帯域について算出し、前記セグメントに属する全ての前記周波数帯域について算出した平均マグニチュードを前記セグメント区間の平均ピッチにおけるターゲットスペクトル・シェイプとして、前記音声変換用辞書に格納する、ことを特徴としている。

【0037】

【発明の実施の形態】次に図面を参照して本発明の好適な実施形態について説明する。

〔A〕 第1実施形態

まず、本発明の第1実施形態について説明する。

〔1〕 音声変換装置の全体構成

図1に実施形態の音声変換装置（音声変換方法）をカラオケ装置に適用し、ものまねを行うことができるカラオケ装置として構成した場合の例である。音声変換装置10は、歌唱者の音声が入力され、歌唱信号を出力する歌唱信号入力部11と、予め定めたコードブックに基づいて歌唱信号から各種特徴ベクトルを抽出する認識特徴分析部12と、歌唱信号のSMS（Spectral Modeling Synthesis）分析を行って入力SMSフレームデータおよび有声／無声情報を出力するSMS分析部13と、各種コードブックおよび各音素の隠れマルコフモデル（HMM）を予め記憶した認識用音素辞書記憶部14と、曲に依存したターゲット挙動データを記憶するターゲット挙動データ記憶部15と、キー情報、テンポ情報、似具合パラメータ、変換パラメータなどの各種パラメータを制御するためのパラメータコントロール部16と、ターゲット挙動データ記憶部に記憶されたターゲット挙動データ、キー情報およびテンポ情報に基づいてデータ変換を行い、変換された持続時間付音素表記情報、ピッチ情報およびアンプリチュード（振幅）情報を生成し出力するデータ変換部17と、を備えて構成されている。

【0038】また、音声変換装置10は、抽出された特徴ベクトル、各音素のHMMおよび持続時間付音素表記情報に基づいて歌唱者が対象としている曲中のどの部分を歌っているかをビタビアルゴリズムを用いて求め、アライメント情報（＝ターゲット歌手が歌うべき曲中の歌唱位置および音素）を検出するアライメント処理部18と、ターゲット歌手に依存するスペクトル・シェイプ情報を記憶するターゲット音素辞書記憶部19と、アライメント情報、ターゲット挙動データのピッチ情報、ターゲット挙動データのアンプリチュード情報、入力SMSフレームデータおよびターゲット音素辞書のスペクトル・シェイプ情報に基づいてターゲットのフレームデータ（以下、ターゲットフレームデータという。）TGFLを生成し出力するターゲット・デコード部20と、パラメータコントロール部16から入力される似具合パラメ

ータ、ターゲットフレームデータTGFLおよびSMSフレームデータFMSLに基づいてモーフィング処理を行い、モーフィングフレームデータMFLを出力するモーフィング処理部21と、モーフィングフレームデータMFLおよびパラメータコントロール部16より入力された変換パラメータに基づいて変換処理を行い、変換フレームデータMMFLを出力する変換処理部22と、を備えて構成されている。

【0039】さらに、音声変換装置10は、変換フレームデータMMFLのSMS合成を行い、変換音声信号である波形信号SWAVを出力するSMS合成部23と、SMS分析部13からの有声／無声情報に基づいて波形信号SWAVあるいは入力された歌唱信号SVのいずれかを選択的に出力する選択部24と、パラメータコントロール部16からのキー情報およびテンポ情報に基づいて音源部25を駆動するシーケンサ26と、選択部24から出力された波形信号SWAVあるいは歌唱信号SVと音源部25からの出力信号であるミュージック信号MSCを加算して出力する加算部27と、加算部27の出力信号を増幅等行ってカラオケ信号として出力する出力部28と、を備えて構成されている。

【0040】ここで、音声変換装置の各部の構成の説明に先立ち、SMS分析について説明する。SMS分析では、まず標本化された音声波形に窓関数を乗じた音声波形(Frame)を切り出し、高速フーリエ変換(FFT)を行って得られる周波数スペクトルから、正弦波成分と残差成分とを抽出する。この場合において、正弦波成分とは、基本周波数(Pitch)および基本周波数の倍数にあたる周波数(倍音)の成分をいう。そして、正弦波成分として本実施形態では、基本周波数、各成分の平均アンプリチュードおよびスペクトル包絡をエンベロープとして保持する。また、残差成分とは、入力信号から正弦波成分を除いた成分であり、本実施形態では周波数領域のデータとして保持する。さらに得られた正弦波成分および残差成分で示される周波数分析データは、フレーム単位で記憶されることとなる。このとき、フレーム間の時間間隔は固定(例えば、5ms)となっているので、フレームをカウントすることによって時間を特定することができる。さらに各フレームには曲の冒頭からの経過時間に相当するタイムスタンプが付されている。

【0041】[2] 音声変換装置の各部の構成

[2.1] 認識用音素辞書記憶部

認識用音素辞書記憶部14は、コードブック及び音素の隠れマルコフモデルを記憶している。記憶しているコードブックは、歌唱信号を各種特徴ベクトル(より具体的には、メルケプストラム、差分メルケプストラム、エネルギー、差分エネルギー、ボイスネス(有声音尤度))にベクトル量子化するために用いられる。また、本音声変換装置においては、アライメント処理を行うために音

声認識の一手法である隠れマルコフモデル(HMM)を用いており、HMMパラメータ(初期状態分布、状態遷移確率行列、観測シンボル確率行列)を各音素(/a/、/i/等)について求めたものが記憶されている。

【0042】[2.2] ターゲット挙動データ記憶部
ターゲット挙動データ記憶部15はターゲット挙動データを記憶しており、このターゲット挙動データは、音声変換を行う曲それぞれに対応した曲依存のデータである。具体的には、対象となる曲を物まねの対象となるターゲット歌手が歌ったものから、ピッチ、アンプリチュードの時間的变化を抽出したもの(なお、これらを静的変化成分、ビブラートの時間的变化成分に分離して抽出しておくと、後処理の自由度がより高くなる)および対象となる曲の歌詞に基づいて歌詞を音素列の並びに置き換えた音素表記に持続時間を含めた持続時間付音素表記が含まれる。例えば、持続時間付音素表記は、音素表記/n//a//k//i/……に対し、各々の持続時間、すなわち、/n/の持続時間、/a/の持続時間、/k/の持続時間、/i/の持続時間、……が含まれる。

【0043】[2.3] ターゲット音素辞書記憶部
ターゲット音素辞書記憶部は、物まね対象となるターゲット歌手の各音素に対応したスペクトル情報であるターゲット音素辞書を記憶しており、ターゲット音素辞書には、何種類かのピッチに対応したスペクトル・シェイプおよびスペクトル補間を行うためのアンカーポイント情報が含まれている。ここで、ターゲット音素辞書記憶部19に記憶されている音声変換用辞書としてのターゲット音素辞書の作成について図2及び図3を参照して説明する。

[2.3.1] ターゲット音素辞書

ターゲット音素辞書は、各音素毎にいくつかのピッチに対応してスペクトル・シェイプと、アンカーポイント情報を有している。図2にターゲット音素辞書の説明図を示す。図2(b)、(c)、(d)は、ある音素におけるピッチf0i+1、f0i、f0i-1にそれぞれ対応するスペクトル・シェイプを示したものであり、一つの音素に対して複数の(上述の例の場合、3個)スペクトル・シェイプがターゲット音素辞書には含まれる。このように複数のピッチに対応したスペクトル・シェイプをターゲット音素辞書として持つ理由は、一般的に同一人物が同一の音素を発声したとしても、ピッチに応じてスペクトル・シェイプの形状は多少変化するものだからである。また、図2(b)、(c)、(d)中、点線は周波数軸上で複数の領域に分ける際の境界線であり、各領域の境界の周波数がアンカーポイントであり、アンカーポイント情報として当該周波数がターゲット音素辞書に含まれている。

【0044】[2.3.2] ターゲット音素辞書の作成

次にターゲット音素辞書の作成について説明する。ま

ず、ターゲット歌手がそれぞれの音素について出しうる一番低いピッチから一番高いピッチまで、連続して発生したものを録音する。より具体的には図2(a)のように、時間とともにピッチをあげていくように発声する。このように録音を行う理由は、より正確なスペクトル・シェイプを算出するためである。すなわち、ある固定ピッチで発生したサンプルから分析して求めたスペクトル・シェイプには、実際に存在するフォルマントが必ずしも現れるとは限らないからである。従って、求めるスペクトル・シェイプに正確にフォルマントが現れるようにするために、あるピッチの前後で同じスペクトル・シェイプとみなせる範囲内の分析結果の全てを用いる必要がある。

【0045】同じスペクトル・シェイプと見なせるピッチの周波数範囲を同じセグメントであるとする、 i 番目のセグメントの中心周波数 f_{0i} は、

【数1】

$$f_{0i} = f_i^{(low)} + \frac{f_i^{(high)} - f_i^{(low)}}{2}$$

ここで、 $f_i^{(low)}$ 、 $f_i^{(high)}$ は、ある音素の i 番目のセグメントの境界のピッチ周波数であり、 $f_i^{(low)}$ が低ピッチ側のピッチ周波数を表し、 $f_i^{(high)}$ が高ピッチ側のピッチ周波数を表す。同じセグメントとみなせるピッチにおけるスペクトル・シェイプの全ての値（周波数及びマグニチュードの組）を一つにまとめる。より具体的には、例えば、図3(a)に示すように、同じセグメントとみなせるピッチにおけるスペクトル・シェイプを同一の周波数軸/マグニチュード軸上にプロットする。次に周波数軸上で周波数範囲 $[0, f_s/2]$ を等間隔（例えば30 [Hz]）に分割する。ここで、 f_s は、サンプリング周波数である。

【0046】このときの分割幅を BW [Hz]、分割数を B （バンド番号 $b \in [0, B-1]$ ）とし、各分割範囲内に含まれる実際の周波数及びマグニチュードの組を (x_n, y_n)

ここで、 $n=0, \dots, N-1$ である。

とすると、当該バンド b の中心周波数 f_b 及び平均マグニチュード M_b は、それぞれ、

【数2】

$$M_b = \frac{1}{2N} \sum_{n=0}^{N-1} (y_{n+1} + y_n)$$

$$f_b = \left(b + \frac{1}{2} \right) \cdot BW$$

と計算される。このようにして求めた (f_b, M_b)

ここで、 $b=0, \dots, B-1$ である。

の組が最終的なあるピッチにおけるスペクトル・シェイプである。

【0047】より具体的には、図3(a)に示した周波数及びマグニチュードの組を用いてスペクトル・シェイプを算出した場合には、図3(c)に示すようにターゲット音素辞書に格納すべき、フォルマントがはっきりと現れた良好なスペクトル・シェイプが得られる。これに対し図3(b)に示すように、同じセグメントとみなすことができないようなピッチにおけるスペクトル・シェイプの全ての値（周波数及びマグニチュードの組）を一つにまとめ、まとめた周波数及びマグニチュードの組を用いてスペクトル・シェイプを算出した場合には、図3(d)に示すように、図3(c)の場合と比較してフォルマントがあまりはっきりしないスペクトル・シェイプが得られることとなる。

【0048】[2.4] ターゲット・デコーダ部

[2.4.1] ターゲット・デコーダ部の構成

図4にターゲット・デコーダ部の構成ブロック図を示す。ターゲット・デコーダ部20は、歌唱者及びターゲット歌唱者のピッチ、アライメントおよび既に処理済みのデコードフレームからデコードされるべきフレームに対応する音素が安定状態にあるかあるいは他の音素に移行する遷移状態にあるかを決定する安定状態/遷移状態決定部31と、スムーズなフレームデータの生成のために既に処理済みのデコードフレームを格納するフレームメモリ部32と、安定状態/遷移状態決定部31における決定結果に基づいてデコードされるべきフレームに対応する音素が安定状態にある場合には現在の音素のスペクトル・シェイプを現在のターゲットのピッチ付近の二つのスペクトル・シェイプから後述のスペクトル補間の方法を用いて第1補間スペクトル・シェイプSS1として生成し、デコードされるべきフレームに対応する音素が遷移状態にある場合には遷移元の音素のスペクトル・シェイプを現在のターゲットのピッチ付近の二つのスペクトル・シェイプから後述のスペクトル補間の方法を用いて第2補間スペクトル・シェイプSS2として生成する第1スペクトル補間部33と、を備えて構成されている。

【0049】また、ターゲット・デコーダ部20は、安定状態/遷移状態決定部31における決定結果に基づいてデコードされるべきフレームに対応する音素が遷移状態にある場合に遷移先の音素のスペクトル・シェイプを現在のターゲットのピッチ付近の二つのスペクトル・シェイプから後述のスペクトル補間の方法を用いて第3補間スペクトル・シェイプSS3として生成する第2スペクトル補間部34と、遷移元の音素及び遷移先の音素並びに歌唱者のピッチ、ターゲット歌唱者のピッチ及びスペクトル・シェイプなどを考慮に入れて遷移元の音素から遷移先の音素に遷移させる場合の遷移のさせかたを規定する遷移関数を発生する遷移関数発生部35と、安定状態/遷移状態決定部31における決定結果に基づいて

デコードされるべきフレームに対応する音素が遷移状態にある場合に遷移関数発生部35において発生された遷移関数並びに第2補間スペクトル・シェイプSS2及び第3補間スペクトル・シェイプSS3の二つのスペクトル・シェイプから後述のスペクトル補間の方法を用いて第4スペクトル・シェイプSS4として生成する第3スペクトル補間部36と、を備えて構成されている。

【0050】さらに、ターゲット・デコーダ部20は、出力されるデコードフレームがよりリアルであるようにターゲットのピッチ及びフレームメモリ部32に格納されている処理済みのデコードフレームに基づいてスペクトル・シェイプの微細構造を時間軸に沿って変化させ（例えば、マグニチュードを時間とともに少しずつ変化させる）、時間的変化が付加されたスペクトル・シェイプSSを出力する時間的変化付加部37と、時間的変化付加部37により時間的変化が付加されたスペクトル・シェイプSSをさらにリアルにするためにターゲットのアンプリチュードに対応させてスペクトル・シェイプSSのスペクトル傾きを補正してターゲットスペクトル・シェイプSSTGとして出力するスペクトル傾き補正部38と、アライメント情報、ターゲットのピッチ及びアンプリチュードに基づいて出力するデコードフレームに対応するターゲットのピッチおよびアンプリチュードを算出するターゲットピッチ／アンプリチュード算出部39と、を備えて構成されている。

【0051】[2.4.2] ターゲット・デコーダ部の詳細動作

ここで、ターゲット・デコーダ部20の詳細動作について説明する。この場合において、よりスムーズなフレームデータの生成の為、ターゲット・デコーダ部20が出力すべきフレームデータ（デコードフレーム；ターゲットスペクトル・シェイプ）はフレームメモリ部に記憶される。ターゲット・デコーダ部20への入力情報としては、歌唱音声の情報（ピッチ、アンプリチュード、スペクトル・シェイプ、アライメント）、ターゲット挙動データ（ピッチ、アンプリチュード、持続時間付音素表記）、ターゲット音素辞書（スペクトル・シェイプ）が含まれている。

【0052】そして、安定状態／遷移状態決定部31は、歌唱者、ターゲット歌手のピッチ、アライメント情報、過去のデコード・フレームからデコードされるべきフレームが安定状態（ある音素からある音素への遷移（変化）途中ではなく、ある音素であることが特定できる状態にあるか否かを決定し、決定結果を第1スペクトル補間部33及び第2スペクトル補間部34に通知する。第1スペクトル補間部33は、安定状態／遷移状態決定部31の通知に基づいて、デコードされるべきフレームが安定状態である場合には、現在の音素のスペクトル・シェイプを現在のターゲットのピッチ付近の2つのスペクトル・シェイプから、後述するスペクトル補間の

方法を用いて補間されたスペクトル・シェイプである第1補間スペクトル・シェイプSS1を算出し時間的変化付加部37に出力する。

【0053】また、第1スペクトル補間部33は、安定状態／遷移状態決定部31の通知に基づいて、デコードされるべきフレームが遷移状態である場合には、遷移元の音素（第1の音素から第2の音素に遷移途中の場合における、第1の音素）のスペクトル・シェイプを現在のターゲットのピッチ付近の2つのスペクトル・シェイプから、後述するスペクトル補間の方法を用いて補間されたスペクトル・シェイプである第2補間スペクトル・シェイプSS2を算出し、第3スペクトル補間部36に出力する。一方、第2スペクトル補間部34は、安定状態／遷移状態決定部31の通知に基づいて、デコードされるべきフレームが遷移状態である場合に、遷移先の音素（第1の音素から第2の音素に遷移途中の場合における、第2の音素）のスペクトル・シェイプを現在のターゲットのピッチ付近の2つのスペクトル・シェイプから、後述するスペクトル補間の方法を用いて補間されたスペクトル・シェイプである第3補間スペクトル・シェイプを算出し、第3スペクトル補間部36に出力する。

【0054】これらの結果、第3スペクトル補間部36は、安定状態／遷移状態決定部31の通知に基づいて、デコードされるべきフレームが遷移状態である場合に、第1補間スペクトル・シェイプおよび第2スペクトル補間処理において算出された第2補間スペクトル・シェイプに基づいて後述するスペクトル補間の方法を用いて補間し、第4スペクトル・シェイプSS4を算出し、時間的変化付加部37に出力する。この第4スペクトル・シェイプSS4は、二つの異なる音素の中間的な音素のスペクトル・シェイプに相当するものとなる。この場合において、第4スペクトル・シェイプSS4を求めるべく補間を行う際には、単純にある時間に亘って対応する領域（その境界点はアンカー・ポイントで示される。）内で線形に補間を行うのではなく、遷移関数発生部35において生成される遷移関数に従ってスペクトル補間を行うことにより、より現実に近いスペクトル補間を行うことができる。

【0055】例えば、遷移関数発生部35は、音素/a/から音素/e/に変化する際には、10フレームかけて対応する領域内（後述するアンカー・ポイント間）のスペクトルを時間的に線形に変化させ、また、音素/a/から音素/u/に変化する際には、5フレームかけて変化するが、ある周波数帯域内（後述するアンカー・ポイント間）のスペクトルについては、線形に変化させ、他の周波数帯域内（後述するアンカー・ポイント間）のスペクトルについては、指数関数的に変化させることにより、自然な音素間の移動をスムーズに実現することができる。このため、遷移関数発生処理においては、音素、ピッチに基づくとともに、歌唱者、ターゲットのピッチや

スペクトル・シェイプ等を考慮に入れて、遷移関数を発生させる。この場合において、後述するようにターゲット音素辞書の中にこれらの情報を含めてしまうように構成することも可能である。次に時間的変化付加部37は、入力された第1補間スペクトル・シェイプSS1または第4補間スペクトル・シェイプSS4に対し、ターゲット・デコーダ部20より出力されるターゲットスペクトル・シェイプ(=デコードフレーム)がより実在するフレームと近似するようにターゲットのピッチおよび過去のデコードフレームに基づいて、スペクトル・シェイプの微細構造を変化させ、時間的変化付加スペクトル・シェイプSSとしてスペクトル傾き補正部38に出力する。

【0056】例えば、スペクトル・シェイプの微細構造としてのマグニチュードを時間的に少しずつ変化させるようにする。スペクトル傾き補正部38は、入力された時間的変化付加スペクトル・シェイプSSとして、出力されるターゲットスペクトル・シェイプ(=デコードフレーム)SSTGがより実在するフレームと近似するようにターゲットのアンプリチュードに応じたスペクトル傾きを有するように補正を行い、補正後のスペクトル・シェイプをターゲットスペクトル・シェイプSSTGとして出力する。スペクトル傾き補正処理としては、出力する音量が大きいときは一般的にスペクトル・シェイプの高域が豊か(リッチ)であり、音量が小さいときはスペクトル・シェイプの高域が乏しい(=こもったような音)ことをシミュレートするために、スペクトル・シェイプの高域部の形状を音量に応じて変化させてやるのである。そして、スペクトル傾き補正して得られるターゲットスペクトル・シェイプSSTGをフレームメモリ部32に格納することとなる。一方、ターゲットピッチ/アンプリチュード算出部39は、出力するターゲットスペクトル・シェイプSSTGに対応するピッチTGP、アンプリチュードTGAを算出し出力する。

【0057】[2.4.3] スペクトル補間処理
ここで、図5を参照してターゲット・デコーダ部のスペクトル補間処理について説明する。

[2.4.3.1] スペクトル補間処理の概要

まず、安定状態/遷移状態決定部31における決定結果に基づいてデコードされるべきフレームに対応する音素が安定状態にある場合には、ターゲットデコーダ部20は、当該音素に対応する二つのスペクトル・シェイプをターゲットの音素辞書から取り出し、また、デコードされるべきフレームに対応する音素が遷移状態にある場合には、遷移元の音素に対応する二つのスペクトル・シェイプをターゲットの音素辞書から取り出す。図5(a)及び図5(b)は、安定状態にある音素あるいは遷移元の音素に対応させてターゲット音素辞書から取り出された二つのスペクトル・シェイプであり、この二つのスペクトル・シェイプのピッチは異なっている。例えば、求

めたいスペクトル・シェイプがピッチ140[Hz]、音素/a/のものだとすると、図5(a)のスペクトル・シェイプは、ピッチ100[Hz]の音素/a/に対応するものであり、図5(b)のスペクトル・シェイプは、ピッチ200[Hz]の音素/a/に対応するものである。すなわち、求めたいスペクトル・シェイプのピッチを挟むような前後のピッチでそれぞれ最も近いピッチを有する二つのスペクトル・シェイプであって、かつ、求めたいスペクトル・シェイプと同一の音素に対応する二つのスペクトル・シェイプを用いる。

【0058】得られた二つのスペクトル・シェイプを第1スペクトル補間部33でスペクトル補間の方法で補間することにより、図5(e)に示すような所望のスペクトル・シェイプ(第1スペクトル・シェイプSS1あるいは第2スペクトル・シェイプSS2に相当)を得る。得られたスペクトル・シェイプは、安定状態/遷移状態決定部31における決定結果に基づいてデコードされるべきフレームに対応する音素が安定状態にある場合には、そのまま得られたスペクトル・シェイプを時間的変化付加部37に出力する。、さらに安定状態/遷移状態決定部31における決定結果に基づいてデコードされるべきフレームに対応する音素が遷移状態にある場合には、遷移先の音素に対応する二つのスペクトル・シェイプをターゲットの音素辞書から取り出す。図5(c)及び図5(d)は、遷移先の音素に対応させてターゲット音素辞書から取り出された二つのスペクトル・シェイプであり、この二つのスペクトル・シェイプのピッチも図5(a)及び図5(b)の場合と同様に異なっている。そして得られた二つのスペクトル・シェイプを第2スペクトル補間部34で補間することにより、図5(f)に示すような所望のスペクトル・シェイプ(第3スペクトル・シェイプSS3に相当)を得る。さらにまた、安定状態/遷移状態決定部31における決定結果に基づいてデコードされるべきフレームに対応する音素が遷移状態にある場合には、図5(e)及び図5(f)に示したスペクトル・シェイプを第3スペクトル補間部36でスペクトル補間の方法で補間することにより、図5(g)に示すような所望のスペクトル・シェイプ(第4スペクトル・シェイプSS4に相当)を得る。

【0059】[2.4.3.2] スペクトル補間手法
ここで、スペクトル補間の手法について詳細に説明する。スペクトル補間を用いる目的は、以下の二つに大別される。

(1) 二つの時間的に連続するフレームのスペクトル・シェイプを補間し、時間的に二つのフレームの間にあるフレームのスペクトル・シェイプを求める。

(2) 二つの異なる音のスペクトル・シェイプを補間し、中間的な音のスペクトル・シェイプを求める。

図6(a)に示すように、補間のもととなる二つのスペクトル・シェイプ(以下、便宜上、第1スペクトル・シ

エイブSS11および第2スペクトル・シェイブSS12とする。なお、これらは、上述の第1スペクトル・シェイブS1および第2スペクトル・シェイブS2とは全く別個のものである。)を各々周波数軸上で複数の領域Z1、Z2、……に分割する。そして、各領域を区切る境界の周波数を各スペクトル・シェイブ毎にそれぞれ以下のように設定する。この設定した境界の周波数をアンカー・ポイントと呼んでいる。

第1スペクトル・シェイブSS11: RB1,1、RB2,1、……、RBN,1

第2スペクトル・シェイブSS12: RB1,2、RB2,2、……、RBM,2

【0060】図6(b)に線形スペクトル補間の説明図を示す。線形スペクトル補間は、補間位置により定義され、補間位置Xは、0から1までの範囲である。この場合において、補間位置X=0は、第1スペクトル・シェイブSS11そのもの、補間位置X=1は第2スペクトル・シェイブSS12そのものに相当する。図6(b)は、補間位置X=0.35の場合である。また、図6(b)において、縦軸上の白丸(○)は、スペクトル・シェイブを構成する周波数およびマグニチュードの組のそれぞれを示す。従って、紙面垂直方向にマグニチュード軸が存在すると思われるのが適当である。補間位置X=0の軸上の第1スペクトル・シェイブSS11の注目するある領域Ziに対応するアンカー・ポイントが、RB1,1およびRB1+1,1

であり、当該領域Ziに属する具体的な周波数およびマグニチュードの組のうちいずれかの組の周波数=f i1であり、マグニチュード=S1(f i1)であるものとする。補間位置X=1の軸上の第2スペクトル・シェイブSS12の注目するある領域Ziに対応するアンカー・ポイントが、

RB1,2およびRB1+1,2

であり、当該領域Ziに属する具体的な周波数およびマグニチュードの組のうちいずれかの組の周波数=f i2で

あり、マグニチュード=S2(f i2)であるものとする。ここで、スペクトル遷移関数f trans1(x)及びスペクトル遷移関数f trans2(x)を求める。

【0061】例えば、これらを最も簡単な線形関数で表すすると、以下のようになる。

$$f_{trans1}(x) = m1 \cdot x + b1$$

$$f_{trans2}(x) = m2 \cdot x + b2$$

ここで、

$$m1 = RB_{i+1,2} - RB_{i,1}$$

$$b1 = RB_{i,1}$$

$$m2 = RB_{i+1,2} - RB_{i+1,1}$$

$$b2 = RB_{i+1,1}$$

である。次に第1スペクトル・シェイブSS11上に実在する周波数およびマグニチュードの組に対応する補間スペクトル・シェイブ上の周波数およびマグニチュードの組を求める。

【0062】まず、第1スペクトル・シェイブSS11上に実在する周波数およびマグニチュードの組、具体的には、周波数f i1、マグニチュードS1(f i1)に対応する第2スペクトル・シェイブ上の周波数=f i1,2、マグニチュード=S2(f i1,2)を以下のように算出する。

【数3】

$$f_{i1,2} = \frac{W_2}{W_1} (f_{i1} - RB_{i,1}) + RB_{i,2}$$

ここで、

$$W1 = RB_{i+1,1} - RB_{i,1}$$

$$W2 = RB_{i+1,2} - RB_{i,2}$$

である。マグニチュード=S2(f i1,2)を算出するにあたり、第2スペクトル・シェイブSS12上に実在する周波数およびマグニチュードの組のうちで周波数=f i1,2をはさむように最も近い周波数をそれぞれ、(+), (-)のサフィックスを付して表すすると、

【数4】

$$S_2(f_{i1,2}) = S_2(f_{i1,2}^{(-)}) + \left(\frac{S_2(f_{i1,2}^{(+)}) - S_2(f_{i1,2}^{(-)})}{f_{i1,2}^{(+)} - f_{i1,2}^{(-)}} \right) \cdot (f_{i1,2} - f_{i1,2}^{(-)})$$

となる。

【0063】以上から、補間位置=xとすると、第1スペクトル・シェイブSS11上に実在する周波数およびマグニチュードの組に対応する補間スペクトル・シェイブ

$$f_{i1,x} = \frac{(f_{trans2}(x) - f_{trans1}(x))}{W_1} (f_{i1} - RB_{i,1}) + f_{trans1}(x)$$

Sx(f i1,x)=S1(f i1)+{S2(f i1,2)-S1(f i1)}・x

同様に、第1スペクトル・シェイブSS11上の全ての周波数およびマグニチュードの組に対して算出す

る。続いて、第2スペクトル・シェイブSS12上に実在する周波数およびマグニチュードの組に対応する補間スペクトル・シェイブ上の周波数およびマグニチュードの組を求める。

【数5】

【0064】まず、第2スペクトル・シェイプSS12上に実在する周波数およびマグニチュードの組、具体的には、周波数 f_{i2} 、マグニチュード $S2(f_{i2})$ に対応する第1スペクトル・シェイプ上の周波数 $=f_{i1,1}$ 、マグニチュード $=S1(f_{i1,1})$ を以下のように算出する。

【数6】

$$f_{i2,1} = \frac{W_1}{W_2} (f_{i2} - RB_{i,2}) + RB_{i,1}$$

$$S1(f_{i2,1}) = S1(f_{i2,1}^{(-)}) + \left(\frac{S1(f_{i2,1}^{(+)}) - S1(f_{i2,1}^{(-)})}{f_{i2,1}^{(+)} - f_{i2,1}^{(-)}} \right) \cdot (f_{i2,1} - f_{i2,1}^{(-)})$$

となる。以上から、補間位置 $=x$ とすると、第2スペクトル・シェイプSS12上に実在する周波数およびマグニチュードの組に対応する補間スペクトル・シェイプ上

ここで、

$$W1 = RB_{i+1,1} - RB_{i,1}$$

$$W2 = RB_{i+1,2} - RB_{i,2}$$

である。マグニチュード $=S1(f_{i1,1,2})$ を算出するにあたり、第1スペクトル・シェイプSS11上に実在する周波数およびマグニチュードの組のうちで周波数 $=f_{i2,1}$ をはさむように最も近い周波数をそれぞれ、 $(+)$ 、 $(-)$ のサフィックスを付して表すとすると、

【数7】

の周波数 $f_{i2,x}$ およびマグニチュード $Sx(f_{i2,x})$ は以下の式で求められる。

【数8】

$$f_{i2,x} = \frac{(f_{trans2}(x) - f_{trans1}(x))}{W_2} (f_{i2} - RB_{i,2}) + f_{trans1}(x)$$

$$Sx(f_{i2,x}) = S2(f_{i2}) + \{S2(f_{i1,2}) - S1(f_{i2})\} \cdot (x - 1)$$

【0065】同様にして、第2スペクトル・シェイプSS12上の全ての周波数およびマグニチュードの組に対して算出する。上述したように第1スペクトル・シェイプSS11上に実在する周波数 f_{i1} およびマグニチュード $S1(f_{i1})$ の組に対応する補間スペクトル・シェイプ上の周波数 $=f_{i1,x}$ 、マグニチュード $=Sx(f_{i1,x})$ 並びに第2スペクトル・シェイプ上に実在する周波数 f_{i2} およびマグニチュード $S2(f_{i2})$ の組に対応する補間スペクトル・シェイプ上の周波数 $f_{i2,x}$ およびマグニチュード $Sx(f_{i2,x})$ の全ての算出結果を周波数順に並び替えることにより、補間スペクトル・シェイプを求める。これらを全ての領域Z1、Z2、……について行い、全周波数帯域の補間スペクトル・シェイプを算出する。上述の例においては、スペクトル遷移関数 $f_{trans1}(x)$ 、 $f_{trans2}(x)$ を線形な関数としたが、二次関数、指数関数など非線形な関数として定義あるいは関数に対応する変化をテーブルとして用意するように構成することも可能である。

【0066】また、アンカー・ポイントに応じてそれらの遷移関数を変更してやることによりより現実に近いスペクトル補間を行うことが可能である。この場合、ターゲット音素辞書の内容は、アンカー・ポイントに付随した遷移関数情報を含めるように構成すればよい。さらに遷移関数情報としては、遷移先の音素に応じて設定するようにすればよい。すなわち、遷移先の音素が音素Bの場合には、遷移関数Yを用い、遷移先の音素が音素Cの場合には、遷移関数Zを用いる等のように設定し、設定

状態を音素辞書に組み込むようにすればよい。さらに歌唱者、ターゲット歌手のピッチやスペクトル・シェイプ等を考慮に入れ、リアルタイムに最適な遷移関数を設定するようにしても良い。

【0067】[3] 全体動作

次に音声変換装置10の全体動作を順を追って説明する。まず、歌唱信号入力部11により、信号入力処理が行われ、歌唱者の歌った信号を入力する。続いて認識特徴分析部12により認識特徴分析処理が行われ、歌唱信号入力部11を介して入力された歌唱信号SVを以降のアライメント処理部18へ入力すべく、認識用音素辞書に含まれるコードブックに基づいてベクトル量子化を行い、各特徴ベクトルVC(メルケプストラム、差分メルケプストラム、エネルギー、差分エネルギー、ボイスネス(有声音尤度)など)を算出する。なお、差分メルケプストラムとは、前フレームと現在のフレームのメルケプストラムの差分値を示す。差分エネルギーとは、前フレームと現在のフレームの信号エネルギーの差分値を示す。ボイスネスとは、ゼロ交差数、ピッチ検出を行うときに求まる検出誤差等から総合的に求められる値、あるいは、総合的に重み付けして求められる値であり、有声音らしさを表す数値である。

【0068】一方、SMS分析部13は、歌唱信号入力部11を介して入力された歌唱信号SVをSMS分析して、SMSフレームデータFSMSを得て、ターゲット・デコード部20およびモーフィング処理部21に出力する。具体的には、ピッチに応じた窓幅で切り出した波形に対して、

(1) 高速フーリエ変換(FFT)処理

- (2) ピーク検出処理
- (3) 有声／無声判別処理およびピッチ検出処理
- (4) ピーク連携処理
- (5) 正弦波成分属性ピッチ、アンプリチュード、スペクトル・シェイプの計算処理
- (6) 残差成分算出処理

が行われる。アライメント処理部18は、認識特徴分析部12により出力された各種特徴ベクトルVC、認識用音素辞書14からの各音素のHMMおよびターゲット挙動データに含まれる持続時間付音素表記情報より、歌唱者が対象としている曲中のどの部分を歌っているかをビタビアルゴリズムを用いて求める。これにより、アライメント情報が求まり、この結果、ターゲット歌手が歌うべきピッチ、アンプリチュード、音素を検出することができる。

【0069】この処理のなかで、歌唱者がある音素をターゲット歌唱者に比較して長く歌った場合には、持続時間付音素表記情報の持続時間を超えてある音素を歌っていると判断し、ループ処理に入る旨をアライメント情報に含めて出力することとなる。これらの結果、ターゲット・デコーダ部20は、アライメント処理部18により出力されたアライメント情報およびターゲット音素辞書19に含まれるスペクトル情報よりターゲット歌手のフレーム情報(ピッチ、アンプリチュード、スペクトル・シェイプ)であるターゲットスペクトル・シェイプSS TG、ピッチTGP、アンプリチュードTGAを算出し、ターゲットフレームデータTGFLとしてモーフィング処理部21に出力する。モーフィング処理部21は、ターゲット・デコーダ部20から出力されたターゲットフレームデータTGFLおよび歌唱信号SVに対応するSMSフレームデータFMS並びにパラメータコントロール部16から入力された似具合パラメータに基づいてモーフィング処理を行い、似具合パラメータに応じた所望のスペクトル・シェイプ、ピッチ、アンプリチュードを有するモーフィングフレームデータMFLを生成し、変換処理部22に出力する。

【0070】変換処理部22は、パラメータコントロール部16からの変換パラメータに従って、モーフィングフレームデータMFLを変形し、変換フレームデータMMFLとしてSMS合成部23に出力する。この場合において、出力アンプリチュードに応じたスペクトル傾き補正を行うことにより、よりリアルな出力音声を得ることが可能となる。また、変換処理部22で行う処理としては、例えば偶数倍音をなくす等の処理があげられる。SMS合成部23は、変換フレームデータMMFLをフレームスペクトルに変換し、逆高速フーリエ変換(IFFT)、オーバーラップ処理および加算処理を行い、波形信号SWAVとして選択部24に出力する。選択部24は、SMS分析部13からの有声／無声情報に基づいて歌唱信号SVに対応する歌唱者の音声が無声音である場

合には、歌唱信号SVをそのまま加算部27に出力し、歌唱信号SVに対応する歌唱者の音声がある場合には、波形信号SWAVを加算部27に出力する。

【0071】これらの動作と並行して、シーケンサ26は、パラメータコントロール部16の制御下で音源25を駆動してミュージック信号SMSCを生成して加算部27に出力する。加算部27は、選択部24から出力された波形信号SWAVあるいは歌唱信号SVと音源25から出力されたミュージック信号SMSCとを適当な割合で混合して加算し、出力部28に出力する。出力部28は、加算部27の出力信号に基づいてカラオケ信号(音声+ミュージック)を出力することとなる。

【0072】[B] 第2実施形態

次に、本発明の第2実施形態について説明する。本第2実施形態が第1実施形態と異なる点は、第1実施形態のターゲット・デコーダ部においては、モーフィング処理部に出力されるスペクトル・シェイプは、ターゲット挙動データに含まれるターゲットのピッチ、アンプリチュードに基づいて算出していたが、本第2実施形態においては、歌唱者のピッチ及びスペクトル傾き情報に基づいて算出している点である。これに伴い、本第2実施形態のSMS分析部では、正弦波成分属性として、ピッチアンプリチュード、スペクトル・シェイプに加えて、スペクトル傾きも算出しておく必要があるが、ターゲット・デコーダ部を除く各部の構成は第1実施形態と同様である。

【0073】[1] ターゲット・デコーダ部

図7に第2実施形態のターゲット・デコーダ部の構成ブロック図を示す。図7において図4の第1実施形態と同様の部分には同一の符号を付し、その詳細な説明を省略する。ターゲット・デコーダ部50は、安定状態／遷移状態決定部31と、フレームメモリ部32と、第1スペクトル補間部33と、第2スペクトル補間部34と、遷移関数発生部35と、第3スペクトル補間部36と、出力されるデコードフレームがよりリアルであるように歌唱者のピッチ及びフレームメモリ部32に格納されている処理済みのデコードフレームに基づいてスペクトル・シェイプの微細構造を時間軸に沿って変化させる(例えば、マグニチュードを時間とともに少しずつ変化させる)時間的変化付加部57と、時間的変化付加部57により時間的変化が付加されたスペクトル・シェイプをさらにリアルにするために歌唱者のスペクトル傾きと既に生成されたスペクトル・シェイプの傾きを比較し、スペクトル・シェイプのスペクトル傾きを補正して補正後のスペクトル・シェイプをターゲットスペクトル・シェイプSSTGとして出力し、フレームメモリ部32にターゲットスペクトル・シェイプSSTGを格納するスペクトル傾き補正部58と、ターゲットピッチ／アンプリチュード算出部39と、を備えて構成されている。

【0074】[2] 第2実施形態の動作

本第2実施形態の動作は全体としては、第1実施形態と同様であるので、主要部の動作のみを説明する。ターゲット・デコーダ部50の時間的变化付加部57は、出力されるデコードフレームであるターゲットフレームがよりリアルであるように歌唱者のピッチ及びフレームメモリ部32に格納されている処理済みのデコードフレームに基づいてスペクトル・シェイプ（第1スペクトル・シェイプSS1あるいは第4スペクトル・シェイプSS4）の微細構造を時間軸に沿って変化させて（例えば、マグニチュードを時間とともに少しずつ変化させて）、スペクトル傾き補正部58に出力する。スペクトル傾き補正部58は、ターゲット・デコーダ部50から出力するターゲットスペクトル・シェイプSSTGをさらにリアルにするために歌唱者のスペクトル傾きと既に生成されたスペクトル・シェイプの傾きを比較し、スペクトル・シェイプのスペクトル傾きを補正して補正後のスペクトル・シェイプをターゲットスペクトル・シェイプSSTGとして出力し、フレームメモリ部32にターゲットスペクトル・シェイプSSTGを格納する。より具体的には、歌唱者のスペクトル傾きと生成されたターゲットのスペクトル・シェイプのスペクトル傾きの差であるスペクトル傾き補正值（Tilt Correction値）を算出し、図8に示すように、スペクトル傾き補正值に応じた特性を有するスペクトル傾き補正フィルタを生成されたターゲットのスペクトル・シェイプに対してかける。これにより、より自然なスペクトル・シェイプを得ることが可能となる。

【0075】[C] 実施形態の変形例

[1] 第1変形例

ピッチ、アンプリチュードに関して、前もって静的変化成分と、ビブラートの静的変化成分（ビブラートを早さ、深さのパラメータとして有する）に分けた情報として持っていれば、例えば、同じ音素を歌唱者がターゲットに比較して長く歌った場合でも、適切なビブラートを付加したピッチ、アンプリチュードを生成することができるので、自然な音の伸びを得ることができる。このような処理を行う理由としては、このような処理を行わない場合には、歌唱者がターゲット歌手と比較して長く音をのばした場合などには、途中でビブラートがかからなくなるなどの現象が生じ、不自然なものとなり、また、歌唱者がターゲット歌手と比較してテンポを変更した場合については、ビブラート成分を持っていない場合には、テンポを上げるとビブラートが早くなってしまい同様に不自然なものとなるからである。

【0076】[2] 第2変形例

以上の説明においては、ターゲット歌唱者の残差成分については、考慮していないものであったが、ターゲット歌唱者の残差成分を考慮する場合に、全てのフレームについて残差成分を保持することは、情報圧縮の観点からいっても本音声変換装置のシステムには適合しない。そ

こで、残差について予め代表的なスペクトルエンベロープを用意し、これらのスペクトルエンベロープを特定するためのインデックス情報を持つようにすればよい。より具体的には、ターゲット挙動データとして残差スペクトルエンベロープ情報インデックスを持たせ、例えば、歌唱経過時間0秒～2秒の間は、残差スペクトルエンベロープ情報インデックス=1のスペクトルエンベロープを使用し、歌唱経過時間2秒～3秒までは残差スペクトルエンベロープ情報インデックス=3のスペクトルエンベロープを使用する。そして、残差スペクトルエンベロープ情報インデックスに対応するスペクトルエンベロープから実際の残差スペクトルを生成して、モーフィング処理において用いるようにすれば、残差についてもモーフィングを可能とすることができる。

【0077】

【発明の効果】本発明によれば、入力された歌唱者の音声を目標とするターゲット歌唱者の歌い方に似せることができるとともに、ターゲット歌唱者の分析データの容量を低減して、リアルタイムに処理を行うことが可能となる。

【図面の簡単な説明】

【図1】実施形態にかかる音声変換装置の概要構成ブロック図である。

【図2】ターゲット音素辞書の説明図（その1）である。

【図3】ターゲット音素辞書の説明図（その2）である。

【図4】第1実施形態のターゲット・デコーダ部の概要構成ブロック図である。

【図5】ターゲット・デコーダ部のスペクトル補間処理の説明図（その1）である。

【図6】ターゲット・デコーダ部のスペクトル補間処理の説明図（その2）である。

【図7】第2実施形態のターゲット・デコーダ部の概要構成ブロック図である。

【図8】第2実施形態のスペクトル傾き補正フィルタの特性説明図である。

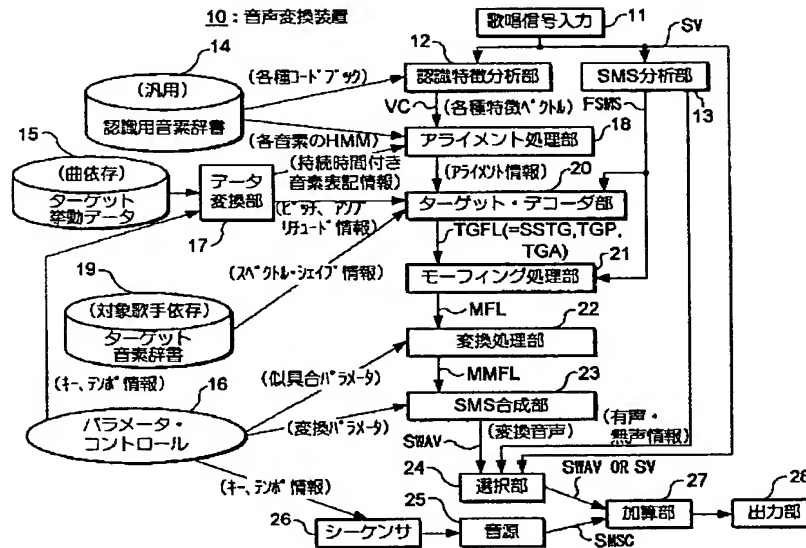
【符号の説明】

10…音声変換装置、11…歌唱信号入力部、12…認識特徴分析部、13…SMS分析部、14…認識用音素辞書、15…ターゲット挙動データ、16…パラメータコントロール部、17…データ変換部、18…アライメント処理部、19…ターゲット音素辞書、20…ターゲット・デコーダ部、21…モーフィング処理部、22…変換処理部、23…SMS合成部、24…選択部、25…音源、26…シーケンサ、27…加算部、28…出力部、31…安定状態／遷移状態決定部、32…フレームメモリ部、33…第1スペクトル補間部、34…第2スペクトル補間部、35…遷移関数発生部、36…第3スペクトル補間部、37…時間的变化付加部、38…スベ

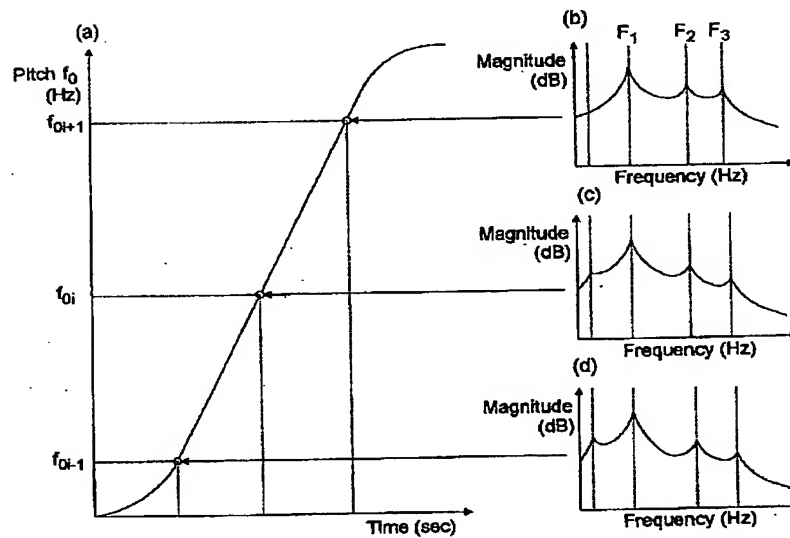
クトル傾き補正部、39…ターゲットピッチ／アンブリ
 チュード算出部、50…ターゲット・デコーダ部、57
 …時間的変化付加部、58…スペクトル傾き補正部、S
 S1…第1スペクトル・シェイプ、SS2…第2スペク

トル・シェイプ、SS3…第3スペクトル・シェイプ、
 SS4…第4スペクトル・シェイプ、SS_t…時間的変
 化付加スペクトル・シェイプ、SSTG…ターゲットス
 ペクトル・シェイプ。

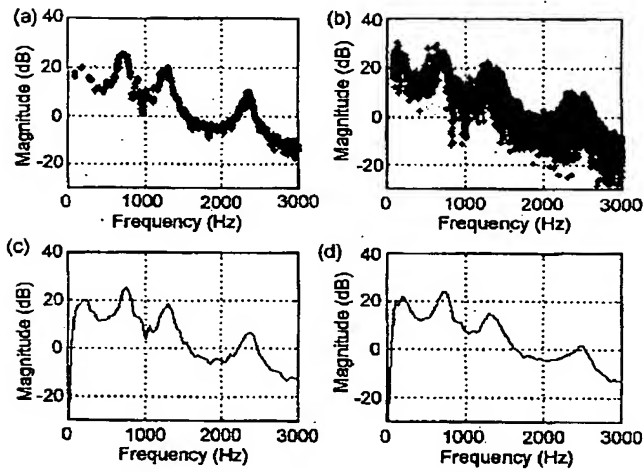
【図1】



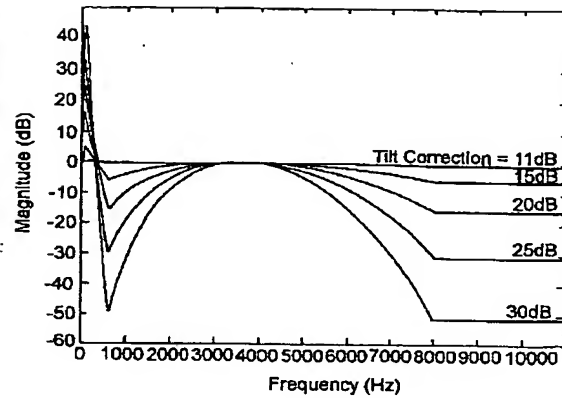
【図2】



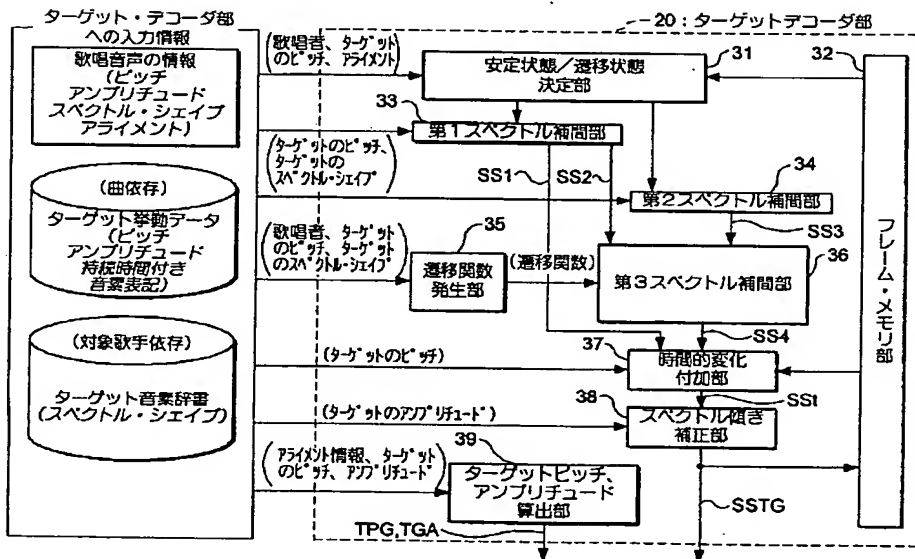
【図3】



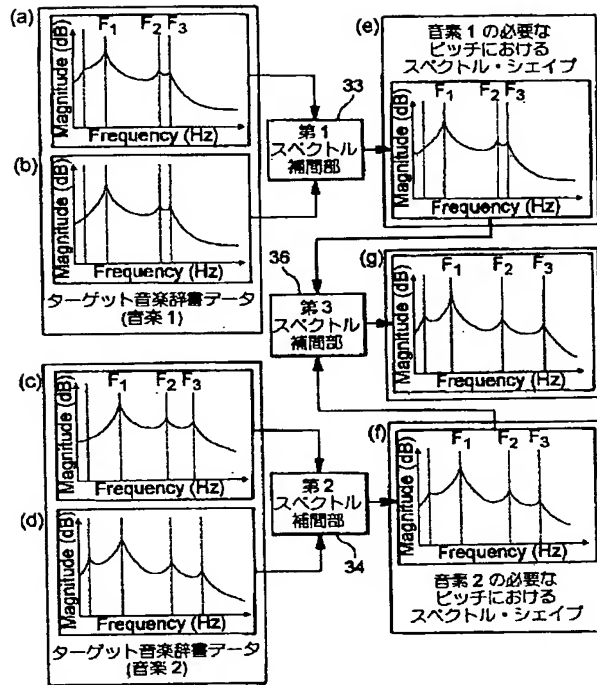
【図8】



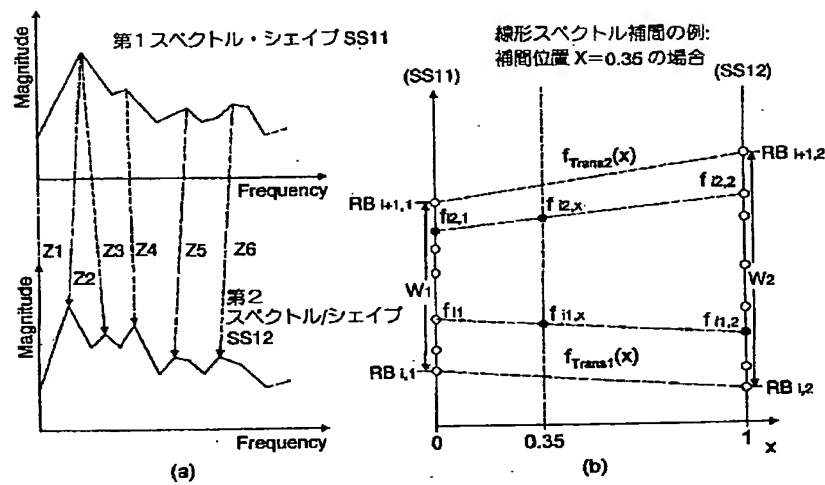
【図4】



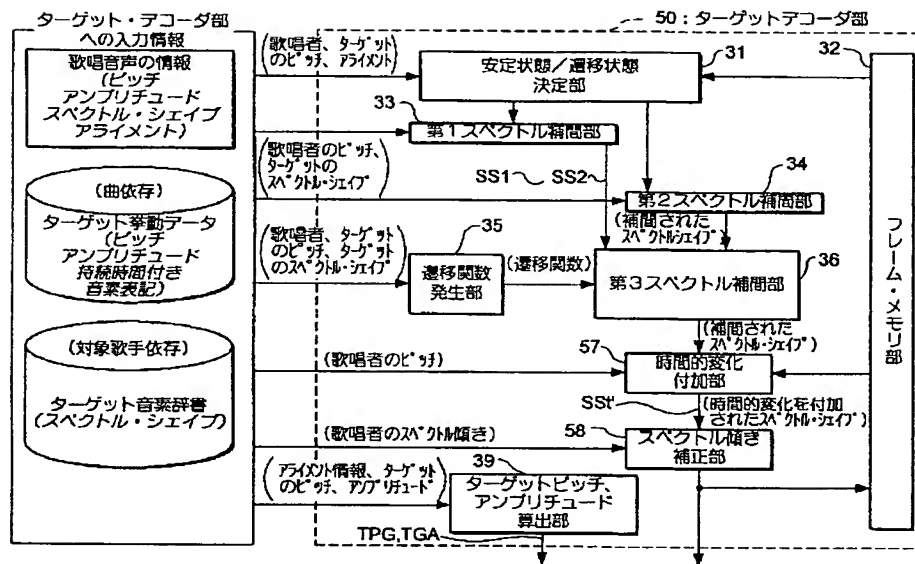
【図5】



【図6】



【図 7】



フロントページの続き

(72)発明者 ザビエル セラ
 スペイン バルセロナ 08002 メルセ
 12

(72)発明者 マーク シーメンツ
 スペイン バルセロナ 08002 メルセ
 12

(72)発明者 ジョルディ ボナダ
 スペイン バルセロナ 08002 メルセ
 12

Fターム(参考) 5D015 AA06 BB02 KK04
 5D045 AA07 BA01
 5D108 BD08 BD11 BF20

THIS PAGE BLANK (USPTO)